Sleep Breath (2013) 17:461–462
DOI 10.1007/s11325-012-0715-1

EDITORIAL (INVITED)

Sleep scoring: man vs. machine?

Christian Berthomier · Marie Brandewinder

Received: 12 March 2012 / Accepted: 25 April 2012 / Published online: 6 May 2012

The automated analysis of sleep has grown in interest in the past decade. Advances in computing have brought the needed intensive calculations within reach [1]; while simultaneously, there is an increasing demand for sleep diagnosis and analysis. The prevalence of sleep troubles is high, and the awareness of their consequences is spreading among patients, health authorities, and clinicians. This awareness is directing more and more patients to sleep centers. The upward trend in demand for sleep evaluations concerns not only sleep specialists. Sleep appears to be an extremely promising territory for other fields, such as cardiology and nutrition for example [2]. Needs exceed capacities by this far. Data analysis has been identified as one of the bottlenecks in the sleep evaluation process, making clear the importance of developing tools to facilitate analysis. These developments have an impact that is medical, as well as economical and social.

The promises of automated sleep analysis are attractive. Its advantages are that it is fast, objective, and reproducible. These qualities may improve data management and patient care. Nothing is more attractive than efficiency and quality. But there are also negative consequences. Dangers include loss of employment for technicians and loss of human expertise. Patient safety may be jeopardized as well. The central issue is, indeed, reliability. First and above all, is automated analysis safe for patients? Is automated analysis useful for clinicians in their routine care of patients? What about clinical trials and research? The controversy is intense. An idea of the entrenched positions can be found in recent issues of *Sleep* [3–5] and the *Journal of Clinical Sleep Medicine* [6]. With all the interests at stake [7], the debate needs carefully conducted studies aimed at evaluating automated software, for instance for automated detection of obstructive sleep apnea [8] or for automated sleep scoring like the one presented by Stege and colleagues in this issue of *Sleep and Breathing*.

While the question—Does it work or not?—is simple, the answer is complicated. To answer this question, there is no simple checkbox for the yes or no. Automated analysis always fails to a degree and succeeds to a certain extent. The comfortable yes/no query should be replaced by a more relative perspective—How much does it work? More correctly put, what is the level of agreement that can and should be expected? This is the potentially controversial question raised by the work of Stege and colleagues.

For automated analysis, the conventional standard is visual scoring. The AASM concentrates a significant part of its activity renewing and clarifying scoring rules and spreading their application through training and center accreditation (http://www.aasmnet.org/ISR/Default.aspx). But as intense as that effort might be, there is and always will be an unavoidable uncertainty with visual scoring. As stated by Silber and colleagues, "no visual-based scoring system will ever be perfect, as all methods are limited by the physiology of the human eye and visual cortex, individual differences in scoring experience, and the ability to detect events viewed using a 30-second epoch" [9]. Studies dealing with interrater variability have reported interscorer agreement of 82 % [10] and shown that the highest agreement, scorers can achieve under very specific conditions, is 87 % [11]. Even for intrascorer agreement, there is nothing like a 100 % agreement. The mean score-rescore agreement is 88 % [11]. Interpreting

C. Berthomier (✉) · M. Brandewinder
Physip,
6, rue Gobert,
75011 Paris, France
e-mail: C.Berthomier@physip.fr

M. Brandewinder
e-mail: M.Brandewinder@physip.fr

the agreement rate between automated scoring and a human sleep expert is, therefore, not straight forward.

Does this mean that, having no clear gold standard, automated analysis can be freed from any stringent performance requirements? Certainly, it is not. The agreement values mentioned above provide a range, which can be seen as the projection on the agreement axis of a scatter plot defined by the experts' scorings. They delineate the acceptable uncertainty. In practice, 85 % of epoch-by-epoch agreement with the consensus of at least two independent scorers seems to be considered as an acceptable value for automated scoring. This value is high in the interscorer agreement range, but that is fair enough.

Such global agreement is not a sufficient criterion though. Indicators that allow the weighting of different sleep stages are also necessary. From a narrow statistical point of view, missing all wake epochs may have little consequence if their proportion is low compared to the sleep period time. Sensitivity and positive predictive value are the statistical measures to compute in order to avoid such misleading interpretations. Specificity and negative predictive value are less discriminating in those circumstances. Likewise, the Pearson correlation coefficient measures the strength of the relation between two variables, scorings here, not the agreement between them [12], and should therefore be avoided and replaced by meaningful Cohen's kappa coefficient [13] and by Bland and Altman plots. Automated analysis assessment is also a methodological challenge.

Criteria of quality may differ from one use to the other, depending upon the focus of the study. As any tool, each automated analysis method can be good for one thing and insufficient for another. Evaluation criteria, as well as conditions of use and limitations, should be made clear.

With clear guidelines for its use, automated analysis can become the valuable help that it is in other domains. Cardiologists, for instance, have a sensible use of automation; it takes care of the analysis and leaves validation and interpretation to the expert. In our case, it can provide extensive information on spectral activity and microstructures and can be an asset for statistical computation to compare conditions, patients, pathologies, sites, and groups.

The automated analysis of sleep has long suffered from a bad image. It is still a work in progress. It needs rigorous and transparent evaluation with elaboration of methods and criteria. Automated analysis also requires regular updates, as opinions can be rapidly obsolete in such a fast moving domain. In sleep, as in any other fields, technology leads to change and should not be a threat when it is constructively developed and carefully managed.

Conflicts of interest Both authors have ownership and directorship in Physip company.

References

1. Penzel T, Hirshkowitz M, Harsh J, Chervin RD, Butkov N, Kryger M, Malow B, Vitiello MV, Silber MH, Kushida CA, Chesson AL Jr (2007) Digital analysis and technical specifications. J Clin Sleep Med 3:109–120
2. Troxel WM, Buysse DJ, Matthews KA, Kip KE, Strollo PJ, Hall M, Drumheller O, Reis SE (2010) Sleep symptoms predict the development of the metabolic syndrome. Sleep 33:1633–1640
3. Svetnik V, Ma J, Soper KA, Doran S, Renger JJ, Deacon S, Koblan KS (2007) Evaluation of automated and semi-automated scoring of polysomnographic recordings from a clinical trial using zolpidem in the treatment of insomnia. Sleep 30:1562–1574
4. Zammit GK (2008) Insufficient evidence for the use of automated and semi-automated scoring of polysomnographic recordings. Sleep 31:449–450
5. Svetnik V, Ma J, Soper K, Doran S, Renger J, Deacon S, Koblan KS (2008) Automated sleep scoring with human supervision adds value compared with human scoring alone: a reply to Zammit G. K. insufficient evidence for the use of automated and semi-automated scoring of polysomnographic recordings. Sleep 31:449–450
6. Schulz H (2008) Rethinking sleep analysis. J Clin Sleep Med 4:99–103
7. Pietzsch JB, Garner A, Cipriano LE, Linehan JH (2011) An integrated health-economic analysis of diagnostic and therapeutic strategies in the treatment of moderate-to-severe obstructive sleep apnea. Sleep 34:695–709
8. Nigro CA, Dibur E, Malnis S, Grandval S, Nogueira F (2012) Validation of ApneaLink Ox™ for the diagnosis of obstructive sleep apnea. Sleep Breath. doi:10.1007/s11325-012-0684-4
9. Silber MH, Ancoli-Israel S, Bonnet MH, Chokroverty S, Grigg-Damberger MM, Hirshkowitz M, Kapen S, Keenan SA, Kryger MH, Penzel T, Pressman MR, Iber C (2007) The visual scoring of sleep. J Clin Sleep Med 3:121–131
10. Danker-Hopfe H, Anderer P, Zeitlhofer J, Boeck M, Dorn H, Gruber G, Heller E, Loretz E, Moser D, Parapatics S, Saletu B, Schmidt A, Dorffner G (2009) Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res 18:74–84
11. Whitney CW, Gottlieb DJ, Redline S, Norman RG, Dodge RR, Shahar E, Surovec S, Nieto FJ (1998) Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 21:749–757
12. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1 (8476):307–310
13. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46